PCT/GB 2004 / 000310

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

PRIORITY DOCUMENT

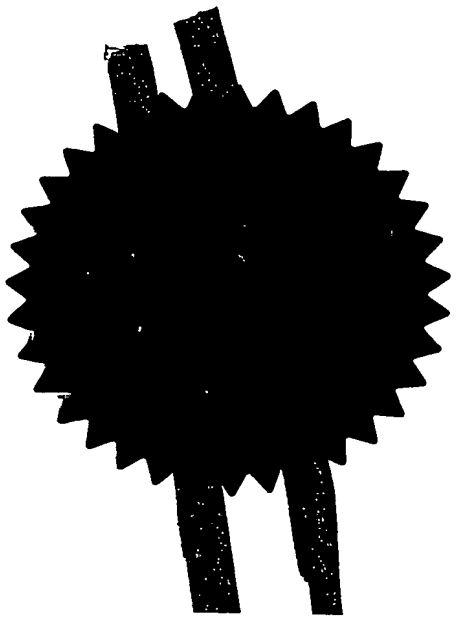
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.



Signed K Behen

Dated 4 March 2004

ents Form 1/77

THE PATENT OFFICE
F
24 JAN 2003
LONDON

The Patent Office

1/77

Patents Act 1977
(Rul

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office
Cardiff Road
Newport
Gwent NP10 8QQ

1. Your reference

A30269

27JAN03 E779901-1 D03052
P01/7700 0.00-0301721.7

2. Patent application number
(The Patent Office will fill in this part)

0301721.7

24 JAN 2003

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

BRITISH TELECOMMUNICATIONS public limited company
81 NEWGATE STREET
LONDON, EC1A 7AJ, England
Registered in England: 1800000

Patents ADP number *(if you know it)*

1867002

If the applicant is a corporate body, give the country/state of its incorporation

UNITED KINGDOM

4. Title of the invention

SEARCH METHOD AND APPARATUS

5. Name of your agent *(if you have one)*

ROBINSON, Simon Benjamin

"Address for Service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

BT GROUP LEGAL SERVICES
INTELLECTUAL PROPERTY DEPARTMENT
HOLBORN CENTRE
120 HOLBORN
LONDON, EC1N 2TE

Patents ADP number *(if you know it)*

1867001

79803110 01

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day/month/year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*

a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.

(See note (d))

YES

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form -

Description 9

Claim(s) -

Abstract -

Drawing(s) 7 &

10. If you are also filing any of the following, state how may against each item

Priority Documents

Translations of priority documents

Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*

Request for preliminary examination and search *(Patents Form 9/77)*

Request for substantive examination *(Patents Form 10/77)*

Any other documents *(please specify)*

11. I/We request the grant of a patent on the basis of this application.

Signature(s) Date: 24 January 2003

ROBINSON, Simon Benjamin , Authorised Signatory

12. Name and daytime telephone number of person to contact in the United Kingdom

Samantha Radley 020 7492 8146

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) *If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.*

b) *Write your answers in capital letters using black ink or you may type them.*

c) *If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.*

d) *If you have answered 'Yes' Patents Form 7/77 will need to be filed.*

e) *Once you have filled in the form you must remember to sign and date it.*

f) *For details of the fee and ways to pay please contact the Patent Office.*

Searching Apparatus and Method

The present invention relates to search engines that access databases. The invention is particularly but not exclusively related to systems that personalises a search engine by creating a user profile.

An example of an application of the invention is to intranet search engines that access large databases such as large corporate repositories holding legal or medical data sets. It also applies to renewed data repositories such as news sources. The invention is typically integrated with a search platform utilised by users who access and search large unstructured databases such as intranets or the Internet. Such platforms may have several thousand users.

Intelligent Personalised Agent Framework, formerly known as Idioms as disclosed in MP Thint, B Crabtree, SJ Soltysiak, Adaptive personal agents, Personal Technologies Journal, 2(3):141-151, 1998; B Crabtree, SJ Soltysiak, Knowing me, knowing you: Practical issues in the personalisation of agent technology, In The PAAM'98 Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, Practical Application Company, March 23-25 1998; and SJ Soltysiak, Intelligent distributed information management systems, Technical report, BTexact Technologies, IS Lab, 1999. This system that acts as a host to a community of users and provides them with on-line services including news sources or corporate databases. The system offers to the users a personalised experience. Within this system, users receive a personalised newspapers everyday using a search engine that has access to an information source such as Intellact disclosed in B Crabtree, SJ Soltysiak, Automatic learning of user profiles - towards personalisation of agent services, BT Technology Journal, 16(3):110-117, 1998.

I Koychev, Tracking changing user interests through prior-leaning of context, In AH'2002, 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, 2002; and D Freitag, J McDermott, D Zabowski, T Mitchel, R Caruana, Experience with a learning personal assistant, Communications of the ACM, 7(37):81 - 91, 1994, disclose profile creation systems that are based on decision tree algorithms that have input vectors with a number of features below thirty. In Koychev's approach the application does not only rely on a window based approach but the algorithm attempts to

freeze an interest in time and save it for future use. When a new interest is found it is checked against "past interests" to see if it corresponds to an old interest, if it does, then the application merges the old interest into the new one; this augments the new interest with information that is relevant to it. The system enables advantageous learning capabilities. Within the scope of Information Retrieval the number of features in a vector are orders of magnitude larger, every keyword that has any relevance must be taken into account and consequently the size of a vector rapidly reaches thousands of features.

In order to adapt user profiles to changes in interests there are two main approaches: the window frame and the ageing mechanism. Maintaining interests in a window frame is a solution that is beneficial to discover and maintain a list of recently introduced interests, because they appear fast and distinctively as shown in Crabtree (1998). However, the drawback of the window frame approach is that it is difficult to retrieve past interests. Typically, if an interest changes or disappears, it is discarded. This has lead to experiments with optimised "interest forgetting functions" as disclosed in I Koychev, Gradual forgetting for adaptation to concept drift, In ECAI 2000 Workshop Current Issues in Spatio-Temporal Reasoning, pages 101 - 106, 2000. This method is a function that decreases the influence of an interest in time; old interests gradually disappear as their importance is reduced linearly over a period of time. The classification of the interests is a crisp set that discards interests when the linear function of the "gradual forgetting" process comes to term.

In order to compensate for the large dimensionality of information retrieval it is known to use user feedback in various forms such as the relevance feedback system disclosed in JJ Rocchio, Performance Indices for Information Retrieval, Prentice Hall, 1971, or user rating as disclosed in D Billsus, M Pazzani, Learning and revising user profiles: The identification of interesting web sites, Machine Learning, 27:313 - 331, 1997. One problem related to requiring feedback from users is that in practice users are reluctant to provide any feedback regardless of how valuable it is to their future requests in the system. It seems that users do not want to interact with the search engine once it has returned the results since it is perceived as an annoyance rather than a benefit.

Embodiments of the invention aim to improve the performance of an on-line search engine by gathering and maintaining user profiles obtained by analysing the documents that are

relevant to the users. The system builds and maintains user profiles in a two fold process. First the system uses an algorithm as disclosed in A Nürnberger, Interactive text retrieval supported by self-organising maps, Technical report, BTexact Technologies, IS Lab, 2002, to extract contextually related keywords from a set of documents. Secondly, the keywords in the concepts are given attributes: life span and a relevance value. The life span indicates to the system when some words within a concept have not been found relevant for some time and therefore should be reduced in importance or removed altogether. The relevance value is a link between two keywords of a concept; this value reflects the strength of the relation between the two keywords. The users have control over these parameters. They can decide if words should have a long or a short life span, and if the strength between keywords should be strong or weak before they can start appearing in their profiles.

The solution proposed here also offers the users the facility to rebuild a query that is more valuable based on their initial query and their profile. The interaction with the system is to be performed before the documents are retrieved when the users are more receptive to more interaction with the system.

This application helps users maintain a profile of temporary interests. The system also provides the analysis required to extract keywords that are relevant to help the users build an efficient profile. The analysis is based on personal data and therefore the keywords suggested to the users are all adapted to their profiles.

The system helps in maintaining profiles, allowing the users to have an informed control over their profile. The system is able to identify which are the keywords and concepts that the users need to improve their search. The profile obtained can be used for query expansion. The users can decide if a keyword is negative or positive to their search.

Embodiments of the invention will now be described with reference to the accompanying figures in which:

figure 1 is a schematic diagram representing the hardware architecture of an embodiment of the invention;

figures 2a and 2b are screen shot of the user interface of an embodiment of the invention showing the embodiment in use;

figure 3 is a schematic illustration of the operation of an embodiment of the invention in response to a user input;

figure 4 is a schematic diagram of the functional elements of the system;

figure 5 is a flow chart illustrating the embodiment of the invention processing data to produce or maintain a list of user interests;

figure 6 is a schematic representation of the processing of the list of interests of figure 5 into a plurality of fuzzy sets.

With reference to figure 1, a conventional PC computer 101 is connected to a network 103 such as a wide area network (WAN) or, more specifically, the Internet. Another computer 105 is connected to the WAN 103 and acts as a server computer. The computers 101, 105 may be connected to the WAN 103 via a Local Area Network (LAN) 107 coupled with the access to a gateway server computer (not shown) that enables the computers 101, 105 to access to the WAN 103. Alternatively, the connection 107 may be provided via home Internet access such as broadband and telephone line based access. The PC computer 101, also referred to as the client machine, is arranged to access the server computer 105. The client machine 101 has software to be able to access the WAN 103. The computer 101 has an operating system (e.g. Microsoft Windows ™, Unix, or Linux) and a web browser (e.g. Microsoft Internet Explorer ™, or Netscape Navigator ™).

An overview of the user interaction with the system will now be described with reference to figures 2a & 2b. On initiation of the system via a web browser the user is presented with a start page 201 as shown in figure 2a. The user can enter a query into the system from a "Search for" box 203 provided. In this example the user enters the acronym for the British Broadcasting Corporation "BBC". A "Search" button 205 instructs the search engine to execute the entered query. In response to this the system returns a list 207 of alternative keywords as shown in figure 2b. In this example the list of keywords 207 comprises the acronyms for some alternative television companies "Granada" and "ITV" as well as the original entry of "BBC". The list of keywords 207 is provided to assist the users perform a better search. The user can select one or more of the keywords from the list 207 to refine their query and then use the "Refine" button 209 to submit the query. The selection can be either positive or negative i.e. the keywords can be included in the query of specifically excluded via alternative selection indicators 211.

As described above, the system returns the list 207 alternative keywords prior to retrieving the search results. Alternatively, the system may be arranged to return the results as would be expected from a conventional search engine. Along with the set of results, the application would return the list 207 of alternative keywords.

The process described above with reference to figures 2a & 2b are summarised in figure 3. The user 301 enters the query into the system 303 at step 305 and system 303 then accesses the user profile 307 for that user at step 309. The system then generates a list of keywords from the profile 307 at step 311 and returns them to the user 301 at step 313 as described above with reference to figure 2b. The user makes their choice of refining the search using the list 207 of keywords and the system executes the query or search at step 315 taking into account the users refinements using the search engine 317 and the database 319. The results are then displayed to the user at step 321 via the system front end.

With reference to figure 4, the core of the system is a profile manager 401 that operates in two phases. The first phase uses a word group extraction system 403 to identify related keywords from a repository of documents 405. The repository 405 is a set of documents that are expected to reflect the users' interests. The extracted groups of related keywords are representative of those interests of a given user. Each user of the system has a document repository 405 which can be maintained either by the user or an automatic document retriever (not shown). The processing of the contents of the repository 405 to extract the related keywords may be performed off-line. The operation of the word group extraction system 403 will be described further below. The second phase is the classification of the related keywords or interests extracted using an interest classifier 407. The interest classifier 407 uses a set of rules 409 to classify interest by their statistical significance (importance) in the corpus of text in the repository 405 and by their age (life span). The operation of the interest classifier 407 will be described further below.

The output of the profile manager 401 is a set of interests 411 classified by their importance in the repository 405 and life span. The profile manager 401 then uses the set of interests 411 in response to the input of a query 413 (203, 205 in figure 2a) to provide the user with a list of keywords (207 in figure 2b). The management and maintenance of the interests is carried out by the profile manager in accordance with a set of rules which

will be described below. The management includes updating the interests from time to time and removing old or outdated interests. The interests 411 are used to refine the search as described above. The set of interests 411 may also be referred to as the user profile. In some situations the profile may include other data describing the users interests and or preferences. The profile manager 401 requires a set of interests 411 before it can provide a list of key words in response to a user query. As a result, the system needs to go through a learning process while the set of interests is initially set up.

The process carried out by the profile manager 401 described above will now be described in further detail with reference to the flow chart of figure 5. At step 501 the profile manager 401 uses the word group extraction system 403 to identify contextually related keywords within bodies of text in the repository 405. The word group extraction system 403 uses a self-organising map (SOM) algorithm disclosed in T Kohonen, Self-organising and associative memory, Springer-Verlag, 1984. The input to the SOM is word triples (represented in a numerical format). The SOM produces a representation of the input words in clusters on a conceptual two-dimensional map where strongly related keywords appear close to one another. For example, if *a*, *b*, *x* and *y* are words that can be found in a text corpus *T*, if the following two word arrangements are frequent across *T*: *a x b*, and *a y b*, then *a* and *b* are contextually related keywords.

At step 503 the output of the SOM algorithm is extracted as a list on contextually related keywords. The list is represented by a number *N* of items made of keywords *A (a,b,c)*, *B (d,e,f)* ... *N (x,y,z)*, where the upper case letters represent sets of related keywords or interests and lower case letters simply represent keywords. The set of interests can be seen as a personalised ontology, every keyword is associated with the keywords that are statistically related to it.

Processing then moves to step 505 at which the profile manager 401 assigns each interest an initial importance value and a life span value. The importance value is initially set up as the average Inverse Document Frequency (IDF) value of every keyword of the interest as disclosed in K Sparck Jones, Index term weighting, Information Storage and Retrieval, (9):313 - 316, 1973. The IDF value of a given keyword reflects its statistical importance into a given text corpus (in this case the user document repository 405). This

importance value is normalised so that the weight can be expressed as a percentage value.

Processing then moves to step 507 where the interest classifier 407 takes each interest in turn and determines whether it is a new interest or an existing interest. If the interest is a new interest processing moves to step 509.

At step 509 if the interest is the first interest for a new set of interests 411 then the profile manager 401 creates a new set and the interest added. If the interest is an addition to an existing set 411 then it is simply added to the set 411.

If at step 507 the new interest is identified as an existing interest in the set 411 then processing moves to step 513. At step 513 each keyword of the new interest is taken in turn and if the keyword is part of the existing interest then its weight is increased by a factor *x*. In the present embodiment the increase is linear and the factor is set to 1.3. If a keyword in the new interest is not present in the existing interest then it is given a weight of 1. Once each keyword in the new interest has been processed in this way the weights are normalised and the system is able to express the weights as a value between 0 and 1.

Step 511 the profile manager 401 gives each interest a life span expressed in days. In the present embodiment this is set to 60 days. A renewed interest is automatically reclassified with a 60 day or full life span. The new or updated interests are then added to the set of interests 411. The existing interest is then replaced with the new or updated interest in the set of interests 401.

Once the profile manager 401 has produced or updated a set of interests 411 it then utilises the interest classifier 407 to process the interests 411 further. With reference to figure 6, the input into the interest classifier is the set of interests 411 and the set of rules 409. The interest classifier 407 outputs the set of interests classified into two fuzzy sets 501, 503. Every interest is classified into one of the three life span fuzzy sets 503a, 503b,

503c and into one of the three importance weight fuzzy sets 501a, 501b, 501c. The classification of each interest depends on the life span and importance weights assigned to each interests in steps 505, 509, 511 and/or 513 of figure 5 as described above.

As noted above, an interest is given an initial life span (step 511 in figure 5) and is classified into one of three fuzzy sets by the interest classifier 407. If the initial classification is "long" the interest will be sustained in the system for at least as long as the system is initially set up to (sixty days in the current implementation). This classification is reviewed on a regular basis by the fuzzy engine such as when concepts are updated or added. If the interest is not renewed its lifespan will result in a gradual downgrading to the "average" set, then to the "short" set and finally will be removed from the set of interests 411. In other words, the classification of an interest into a life span fuzzy set is an indication of its life span expectancy in the system.

The users may have access to the fuzzy sets configuration through an interface to enable them to control the classification process. The users can modify the size of the life span sets 503a, 503b, 503c and thus modify the life span of concepts. To keep concepts longer the fuzzy set of recent concepts 503a is be increased and the sizes of one or more of the sets of older concepts 503b, 503c reduced.

The importance fuzzy sets 501a, 501b, 501c are used in the selection of keywords that will be suggested to a user in response to the entry of a query. For example, the system may be arranged to suggest only strong interests, strong and medium interest or all interests. Again the users can decide on the size of these data sets so that they have control over selection process. Similarly the system 401 is arranged so that if the system is about to discard a concept with strong relevance (because its life span has expired) the system can require confirmation from the user. This gives the user the facility to renew the lifespan of the interest if they choose.

Interests that have had their importance value renewed (step 513 of figure 5) may well remain in the same fuzzy set or they may be upgraded. Others that have not been

renewed may either be sustained a little longer in the same set or they may be downgraded. An interest with an updated importance value is not automatically reclassified in the "high" fuzzy set, others are gradually downgraded to the "medium" and the "low" sets.

The system is designed to help the users manage their profile efficiently. Yet, the system can run without having the users to maintain anything. Users are also allowed to add, change, and remove concepts. They can thoroughly control their sets of interests 411, repositories 405 and rules 409. The system provides a non-obtrusive software application. The application gradually builds fuzzy sets of keywords and is able to make helpful suggestions to the users. By giving control to the users with regards to the size of the fuzzy sets they can manage the maintenance of the profiles and they can build more efficient queries.

Self organising maps are discussed further in T Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics, 43:59-69, 1982; and H Ritter, and T Kohonen, Self-organising semantic maps, Biological Cybernetics, 61(4):241 - 254, 1989.

It will be understood by those skilled in the art that the apparatus that embodies the invention could be a general purpose device having software arranged to provide the an embodiment of the invention. The device could be a single device or a group of devices and the software could be a single program or a set of programs. Furthermore, any or all of the software used to implement the invention can be contained on various transmission and/or storage mediums such as a floppy disc, CD-ROM, or magnetic tape so that the program can be loaded onto one or more general purpose devices or could be downloaded over a network using a suitable transmission medium.

Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise", "comprising" and the like are to be construed in an inclusive as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".

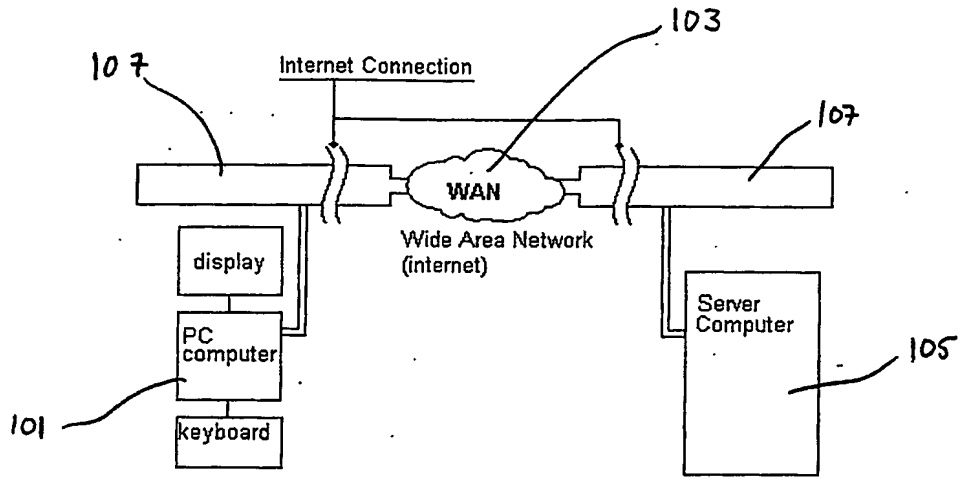

103
107
Internet Connection
107
WAN
Wide Area Network
(internet)
display
PC
computer
keyboard
Server
Computer
105
101

Figure 1

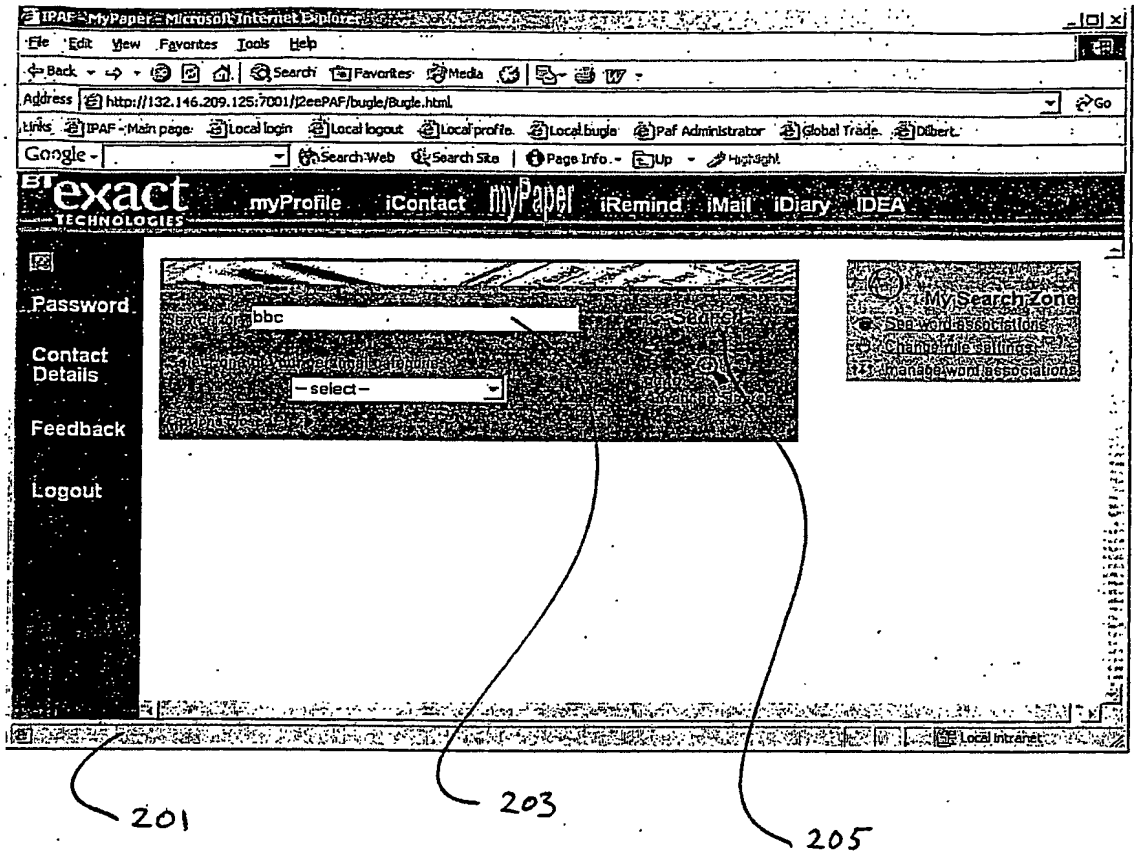

IPAF - MyPaper - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://132.146.209.125:7001/J2eePAF/bugle/Bugle.html
Links IPAF - Main page Local login Local logout Local profile Local bugle Paf Administrator Global Trade Dilbert
BT exact TECHNOLOGIES
myProfile iContact myPaper iRemind iMail iDiary IDEA
Password
Contact Details
Feedback
Logout
bbc
Search
-select-
My Search Zone
Local intranet
201
203
205

Figure 2a

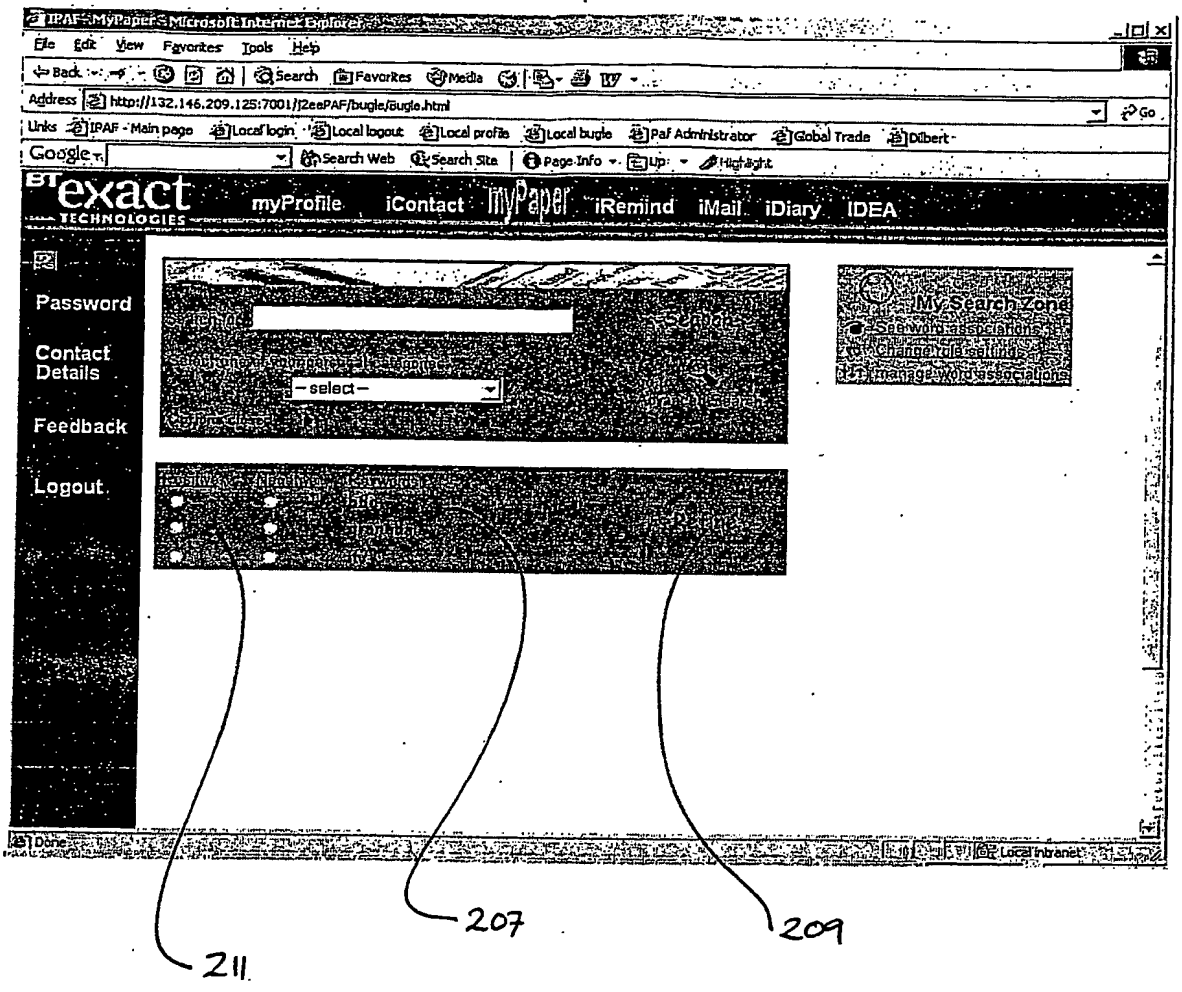
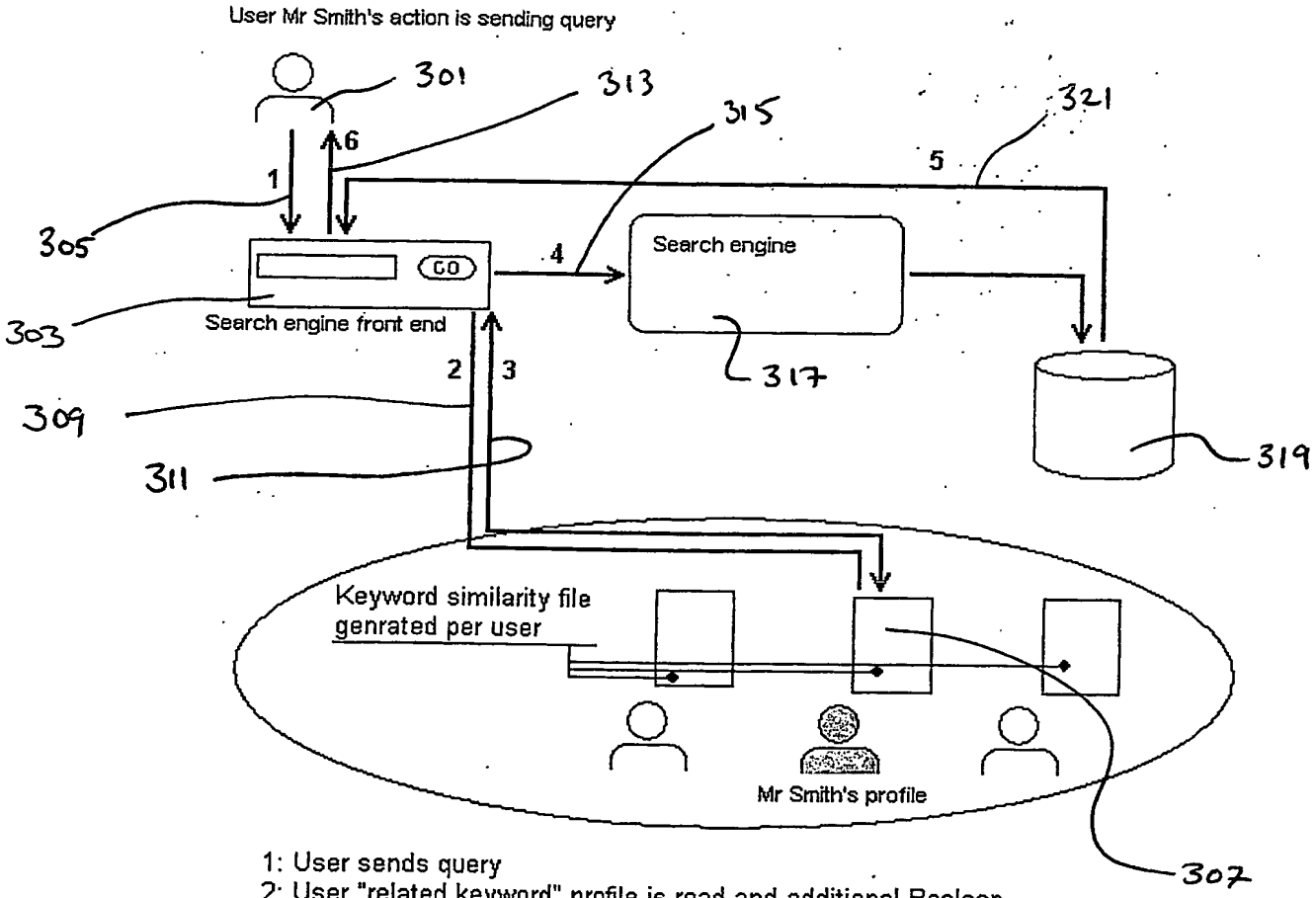
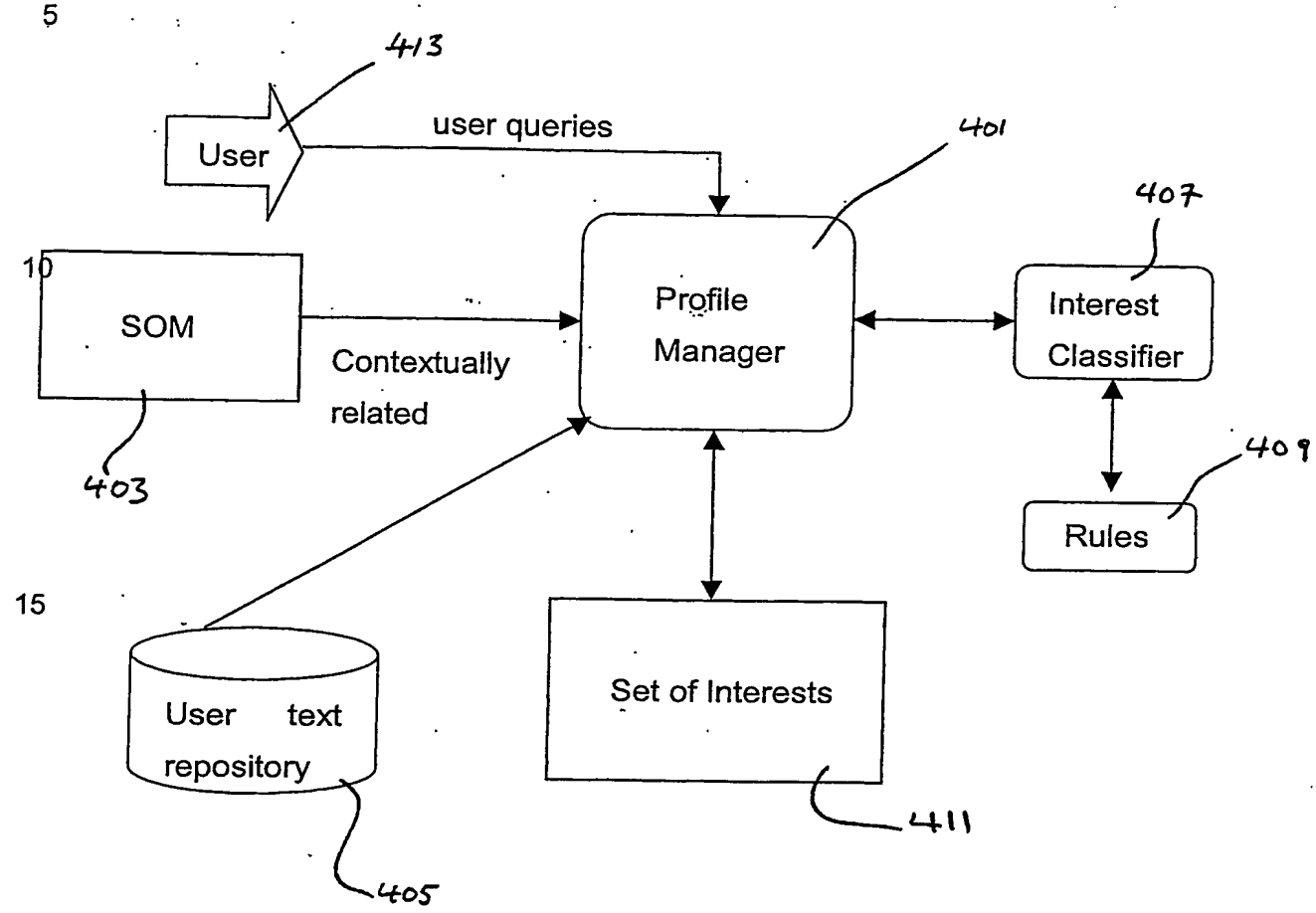

IPAF - MyPaper - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://132.146.209.125:7001/j2eePAF/bugle/Bugle.html
Links IPAF - Main page Local login Local logout Local profile Local bugle Paf Administrator Global Trade Dilbert
BT exact TECHNOLOGIES
myProfile iContact myPaper iRemind iMail iDiary IDEA
Password
Contact Details
Feedback
Logout
-select-
My Search Zone
Done
Local Intranet
211
207
209

Figure 2b



1: User sends query
2: User "related keyword" profile is read and additional Boolean keywords are extracted
3: The user is given the Boolean terms to expand his query
4: The expanded query can go to a search engine (one that can handle Boolean querying)
5: The set of result is extracted from the data repository
6: The Graphical User Interface displays the results back to the user

Figure 3


413
User
user queries
401
407
10
SOM
Profile Manager
Interest Classifier
Contextually related
403
409
Rules
15
User text repository
Set of Interests
411
405

20

Figure 4

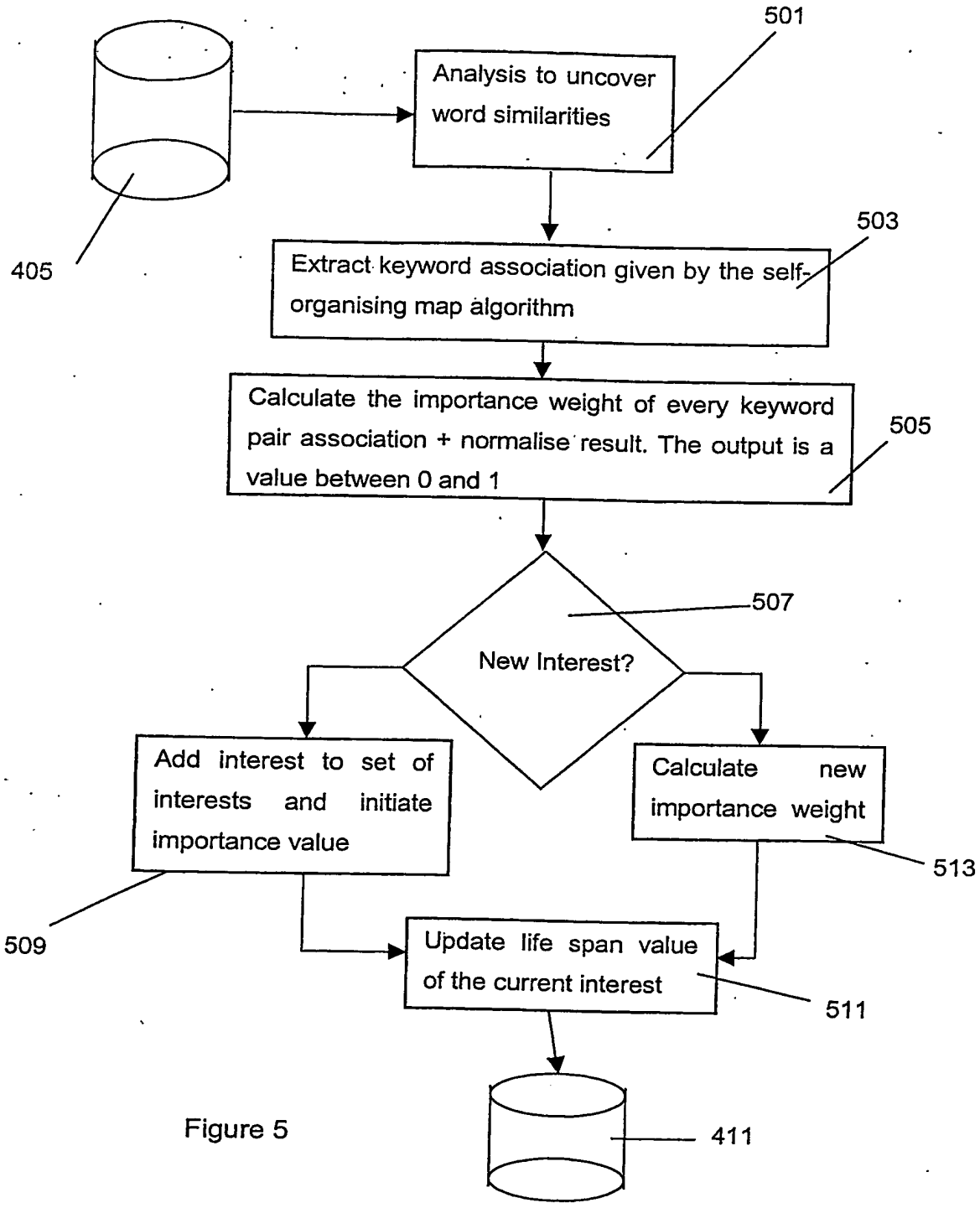

501
Analysis to uncover word similarities
405
503
Extract keyword association given by the self-organising map algorithm
Calculate the importance weight of every keyword pair association + normalise result. The output is a value between 0 and 1
505
507
New Interest?
Add interest to set of interests and initiate importance value
Calculate new importance weight
513
509
Update life span value of the current interest
511
411

Figure 5

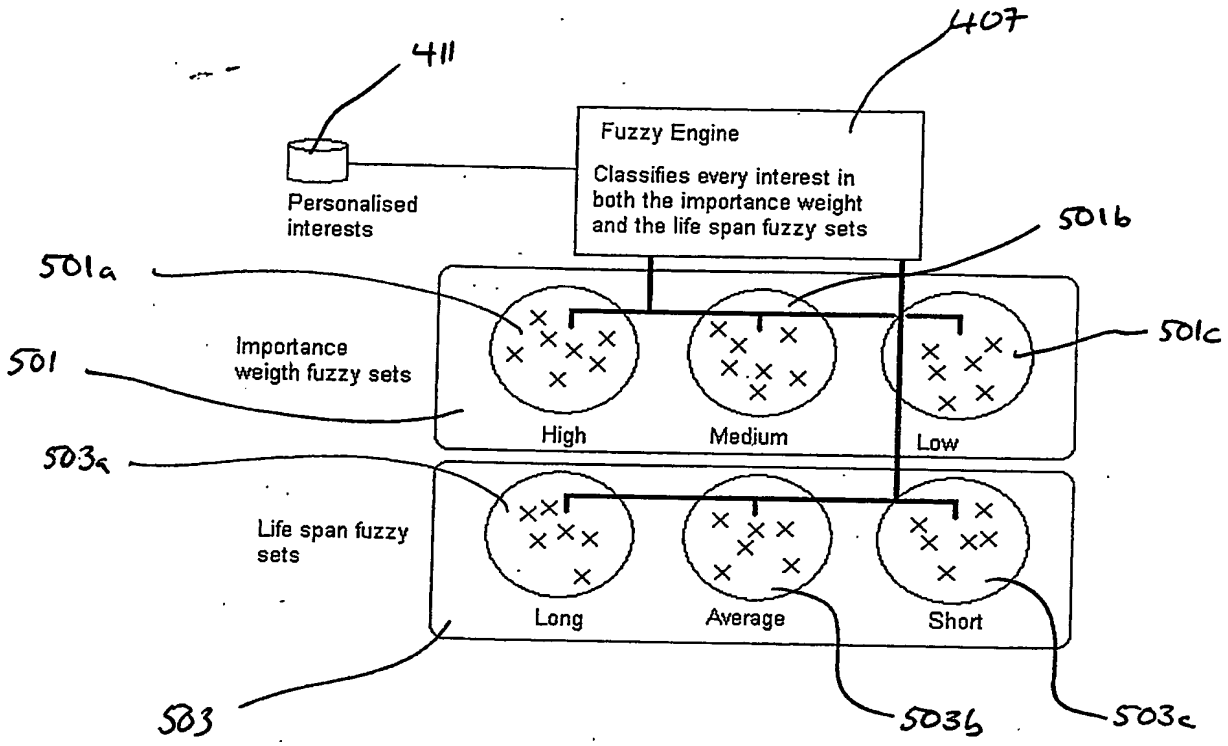

411
407
Fuzzy Engine
Classifies every interest in both the importance weight and the life span fuzzy sets
Personalised interests
501a
501b
501c
501
Importance weigth fuzzy sets
High
Medium
Low
503a
Life span fuzzy sets
Long
Average
Short
503
503b
503c

Figure 6

PCT Application
PCT/GB2004/000310

This Page is inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☑ COLORED OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REPERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning documents *will not* correct images problems checked, please do not report the problems to the IFW Image Problem Mailbox